안녕하세요. 유플러스 AI 대회 싹쓰리팀의 김성연 입니다. 발표 시작하겠습니다.

목차로는 데이터를 다양한 관점에서 바라보는 EDA, 여러가지 모델 접근법, 모델을 실제로 서빙 하는 부분, 마지막으로 모델 개선방안 순서로 진행하겠습니다.

저희 팀은 저, 구진범, 이환주로 구성되어있구요. 각자 자기의 관점에서 EDA를 먼저 진행했고 이를 바탕으로 의견을 서로 공유하고 (라이트GCN, CatBoost, 룰베이스 모델을 나눠서 분담했습니다.)

다음으로 데이터 EDA 부분입니다. 성별을 기준으로 인기 컨텐츠을 나열해 보았는데요. 남자아이와 여자아이의 성향이 생각보다 크게 다르지 않음을 알 수 있습니다.

이번에는 나이를 기준으로 보았습니다. (나이가 들수록 컨텐츠 표지부터 점점 복잡해지는 것을 알 수 있습니다.) 앞서 본 성별보다 나이가 컨텐츠를 선택하는데 중요한 요소로 보입니다.

다음으로 3~4월 상품을 4등분 하여 살펴보았습니다. (꾸준히 인기 있는 상품도 보이고 반짝 인기있는 상품도 관찰되는 것 같아요.)

변수별로 그룹 유사도를 시각적으로 확인했는데요. 리콜과 엔디시지를 이용하여 객관적인 수치로도 확인하였습니다. 성별보단 나이 간 그룹 유사도가 떨어지는 것으로 보아 나이가 유저의 특징을 잘 나타내는 변수라고 할 수 있습니다. 또 날짜도 기간 차이가 많이 날수록 그룹 유사도가 떨어지는 것을 확인할 수 있습니다.

부모나 아이의 관심 키워드도 한번 살펴봤는데요. 노래/율동 그룹이 타 그룹 대비 이질적인 컨텐츠 선호를 보입니다.

다음으로 전반적인 데이터 구조에 대해 알아보겠습니다. 상위 1% 아이템이 전체 데이터의 26%을 나타내고 있고 5명 이하로 시청한 컨텐츠는 거의 60%가 됩니다. 인기있는 컨텐츠에 상당한 편향이 있어 질 좋은 추천 모델을 만들기 까다롭습니다.

다음으로 유저와 컨텐츠 수 대비 데이터의 개수가 어느정도인지 알아보았습니다. 약 0.2% 데이터가 유효한데요. 이 정도는 추천 모델을 적용하기에 충분합니다.

Train 데이터는 3월~4월, test 데이터는 5~7월 입니다. Train 데이터 기간이 짧지만 최대한 test 데이터와 유사한 환경을 갖기 위해 시계열 성질을 고려하여 4월 15일 이후를 valid 데이터로 설정했습니다. 해당 작업을 수행하니 valid scoe와 test score가 꽤 비례하는 모습을 보였습니다.

기존 baseline과 대비해 데이터 내 다양한 컨텐츠 정보를 추가해준 catboost 모델을 사용해보았습니다. 범주형 변수가 많기 때문에 이 부분에 최적화된 catboost 모델 학습을 진행했습니다.

각 유저의 아이템 별 평균 시청 시간 비율(아이템의 run time 대비 시청 시간)을 타겟으로

/ 나이, 관심 키워드, sub title, cast를 feature로 사용하였을 때 가장 높은 성능을 보여주었습니다.

앞서 말씀드린 train과 valid를 바탕으로 분석한 결과 각각의 유저들은 신규 컨텐츠를 찾는 것 보다 기존 시청 컨텐츠를 소모하는 경우가 대부분입니다. 대회 평가 기준인 리콜과 엔디시지를 위해선 유저별 기존 시청 컨텐츠를 먼저 추천해준 다음에 켓부스트와 같은 모델에서 나온 결과를 이어서 추천해주는 방식을 사용했습니다. 그렇기 때문에 기존 시청 컨텐츠가 적은 유저에게 뒤에 붙이는 모델이 더 중요하게 작동될 것 입니다.

이어서 추천해주는 방식 중 단순하지만 강력했던 방법이 나이 군집 별로 인기 상품을 추천해주는 것 입니다. 이 방식이 단순하기 때문에 다른 머신러닝 모델들이 효과적인지 판단하는 기준점이 될 수 있습니다. 실제로 컨텐츠 인기도의 편향이 심해서인지 나이 군집 별 인기 상품 추천이 상당히 강력했습니다.

다음으로 기존 베이스라인에서 나온 추천 결과를 사용해보았는데요. 이때 하이퍼파라미터 튜닝은 기존 시청 컨텐츠에 이어서 붙인 valid score가 높은지를 고려하였습니다. 또한 추출된 임베딩으로 군집을 만들고 군집 별 인기 상품을 추천해보기도 하였습니다.

마지막으로 유저-아이템간 상호작용을 잘 고려하는 그래프 기반 모델을 사용해 군집화를 해보았는데요. 성능이 잘 나오지 않는 것으로 보아 유저-아이템간 상호작용이 깊지 않다는 결론을 내렸습니다.

복잡한 딥러닝 모델이 압도적인 성능을 거두지 못하였기 때문에 저희가 서빙을 한다면 간단한 나이 군집 별 인기 상품을 추천하는 방식으로 진행할 것 입니다.

이 방식은 시퀀스를 고려하지 않으며, 최신 반응을 보고 실시간으로 추천하는 모델이 아니므로 자원 소모가 덜한 배치 서빙으로도 충분합니다.

상황에 따라 서빙 주기는 달라질 수 있겠지만 1일 단위로 해도 충분하다고 생각합니다. 배치 서빙은 실시간 처리가 아니므로 비용 측면에서도 부담이 없고 온라인 서빙보다 구현이 쉽고 간단합니다. 또한, 한 번에 많은 데이터를 처리하므로 latency 문제가 없다는 장점이 있습니다.

코드를 함수화하여 주기적으로 스케줄링 작업을 진행할 때 사용할 수 있는 것으로는 에어플로우와 크론잡 등이 있습니다.

다음으로 저희가 생각하는 추천 모델 개선 방안 입니다. 질 좋은 추천 시스템이 리콜과 엔디시지를 극대화 하는 것과 완전히 동일하지는 않다고 생각합니다.

저희가 생각하는 질 좋은 추천시스템을 판단하는 또 다른 요소로는 참신함과 컨텐츠 age, 그리고 얼마나 개별화된 추천을 하는지 입니다. “참신함”을 일부분 고려한다면 인기가 없지만 질 좋은 영상을 가지고 광고처럼 활용하여 수익을 창출할 수 있고 특정 인기 컨텐츠만 소비되지 않는 건강한 컨텐츠 생태계를 만들 수 있습니다. 또 오래된 영상보다 신규 컨텐츠에 가중치를 두는 것도 좋은 방식이 됩니다.

다음으로 모델을 고도화 하기 위해선 추천을 진행한 후에 클릭 여부를 나타내는 피드백 로그를 생성한다면 negative sampling을 하지 않는 완성도 높은 모델을 만들 수 있습니다. 또 아이템의 내용 정보가 조금 부족하다고 느껴졌는데 내용 정보를 더 많이 수집함으로써 컨텐츠 기반으로 최신 컨텐츠를 더 잘할 추천할 수 있습니다.

그리고 앞서 EDA에서 살펴봤듯이 시간에 따라 인기 아이템이 금방 바뀌므로 모델 최신화를 주기적으로 진행해야합니다.

마지막으로 연관성과 관련된 지표인 recall이나 NDCG 외에 참신함이나 새로움과 같은 지표도 복합적으로 고려하면 더 좋은 추천 모델을 완성할 수 있다고 생각합니다.